Classification of Twitter Accounts into Automated Agents and Human Users

Zafar Gilani, Ekaterina Kochmar, Jon Crowcroft
Computer Laboratory
University of Cambridge
{szuhg2, ek358, jac22}@cam.ac.uk

***Abstract*—Online social networks (OSNs) have seen a remarkable rise in the presence of surreptitious automated accounts. Massive human user-base and business-supportive operating model of social networks (such as Twitter) facilitates the creation of automated agents. In this paper we outline a systematic methodology and train a classifier to categorise Twitter accounts into 'automated' and 'human' users. To improve classification accuracy we employ a set of novel steps. First, we divide the dataset into four popularity bands to compensate for differences in types of accounts. Second, we create a large ground truth dataset using human annotations and extract relevant features from raw tweets. To judge accuracy of the procedure we calculate agreement among human annotators as well as with a bot detection research tool. We then apply a Random Forests classifier that achieves an accuracy close to human agreement. Finally, as a concluding step we perform tests to measure the efficacy of our results.**

***Index Terms*—social network analysis; account classification; automated agents; bot detection**

I. Introduction

Twitter, with its 313 million active monthly users, sustains an increasingly large population of automated programs on its platform. This is largely due to its inherently open nature, convenient sign-ups, its 140 character limit per message, and an extensive API that offers a multitude of functionalities to the programming community.

Automated agents, more notoriously referred to as *bots*, are on a sharp rise – on the Web in general and on the social networks in particular [7]. In fact, 51.8% of all Web traffic is thought to be generated by bots.[1] A media analytics company found that 54% of the online ads shown in 2012 and 2013 were viewed by bots rather than humans.[2] In 2014 Twitter itself reported that 13.5 million (5% of the total at the time) of its accounts were either fake, fraudulent or spam.[3] Due to their phenomenal rise and concealed *modus operandi*, bots have usually been associated with malevolent purposes and negative activities. These include posting ads and spamming, link farming,[4] smear campaigning,[5] spreading malicious content or false information, and more recently political infiltration.[6]

Despite this 'negative' rise, not all bots are created exclusively for malevolent purposes. There are bots which are benign and benevolent, such as news and emergency communication, art and discovery,[7] content aggregation, fun and humour,[8] marketing and business promotion, and social activism [16]. Therefore, we introduce the term *agent* to represent both good and bad bots on Twitter, in this paper. To accurately refer to agents in previous research, we will use the term 'bots'.

The existence of these agents is further evident from recent examples. Microsoft's Tay was a bot operating a Twitter account learning to mimic human speech patterns by interacting with other users through tweets and replies. The experiment had to be terminated when Tay was taught hate-speech and racism.[9] This highlights that automated conversation and content dissemination may take an unexpected turn that the users may find offensive and harmful. Recently, an MIT scientist programmed a Twitter bot that tweets like the US president Donald Trump.[10] The bot uses an AI algorithm to learn Trump's style of speech by going through debate transcripts. This exemplifies the other side of the coin – the recent research trend of automating content generation and mimicking people on Twitter.

In this paper we present a methodology and a mechanism for *non-partisan* classification of Twitter users into automated agents and human users, by refining preprocessing and partitioning of datasets, creating and using a large human annotated dataset as ground truth labels, as well as extracting most relevant feature-sets (via ablation tests) for each popularity band.

II. Motivation and Contributions

The goal of this research is to classify Twitter users as automated agents (that tweet via a scheduling tool or an automated program that uses Twitter API) and human users. We note that we clearly distinguish between our task of agent classification and spam detection. Spam is usually subversive and malicious in nature [18], is often found to be high in volume and frequency, and contains URLs (that point to malicious websites) and spam words [1], [14]. However, automation

[1]Bot traffic report 2016 – http://bit.ly/2kzZ6Nn
[2]Fake ads traffic – http://bit.ly/2cXhfBv
[3]Twitter's 2014 Q2 SEC filing – http://bit.ly/1kBx4M8
[4]Link farming – http://bit.ly/2cXhfBv
[5]Smear campaign – http://bit.ly/2kvYXI8
[6]Social bots distort U.S. election – http://bit.ly/2l3VzGf
[7]Art and discovery bots – http://bit.ly/2lcmPPX
[8]Fun and humour bots – http://bit.ly/2kCu4Ec
[9]Microsoft's Tay – http://bit.ly/2bWnRKV
[10]DeepDrumpf – http://read.bi/2dpchdd

is not exclusively employed for malevolent purposes. There could be many variants of automation due to the usage of APIs and third-party services, and it can often involve direct human intervention (see §I, IV). Also, there are no guarantees that a successfully detected spam account is operated by an agent and not a human – it could be either. This forms a strong basis for detecting automation without any prior judgement.

An implemented research tool that offers an API is BOTORNOT [5]. BOTORNOT uses six feature-sets and a Random Forests classifier to output bot-likelihood score of a given Twitter account. We carry out a well-defined human annotation task (see §IV) and compare these to the BOTORNOT annotations. In our experiments, we have found that BOTORNOT produces an average agreement of 48% with human annotators, while the average agreement among human annotators is 89%.

Our work has the following contributions: (*i*) Use of raw historical data (60 million tweets) for attribute collection and account classification ($722,109$ tweets) to cater for stealthier agents that are harder to discern from humans; (*ii*) A Twitter dataset divided into user popularity bands, further partitioned into lists of agents and humans (for reasons refer to §IV) using a human annotation task. This serves as a large ground truth dataset; (*iii*) 14 novel features from a total feature-set of 21 attributes (see §IV); (*iv*) Performance evaluation of current state of the art in bot detection by calculating agreement between human annotators and BOTORNOT; (*v*) Application of supervised learning approach – Random Forests classifier – for *non-partisan* account categorisation; (*vi*) Identification of a distinct group of features (using ablation tests) that are most informative for classifying automated agents within each popularity band (*cf.* Table VIII); and (*vii*) Hypotheses (*cf.* Table I) verification against our findings using t-tests (see §VI).

III. RELATED WORK

Research has focussed on a number of different aspects of social media. Relevant work can be categorised into four domains: *user behaviour*, *social media infiltration experiments*, *social impact of bots*, and the problem of *bot detection on social networks*. Coincidentally, there has been a recent surge in research focused on automating content generation [17] that looks to have been produced by humans.

User behaviour. In [12] authors used follower-to-following ratio on Twitter to classify the users into broadcasters (having significantly more followers than following), acquaintances (congruent follower-to-following ratio), and miscreants and evangelists. In a related work [20] authors use principal component analysis to identify deviations in anomalous user behaviour from normal user behaviour. The authors then apply unsupervised anomaly detection technique to address the problem of detecting subversive promotion techniques via fake and compromised accounts, and collusion networks or bot farms on Facebook. Both of these works perform user classification to detect subversive and attacker strategies in online social settings, but do not focus on automation.

Social media infiltration experiments. In [2], Boshmaf *et al.* evaluate vulnerability of Facebook against large-scale infiltration by deploying a social bot network of 102 profiles. They found that 86% of bots infiltrated up to 50 user profiles and 10% bots were able to infiltrate up to 80 user profiles. They found that a successful infiltration reveals users' private information, and security defences are not sufficient to guard from a stealthy infiltrator. Similarly, in [8] Freitas *et al.* evaluate infiltration strategies on Twitter using 120 social bot profiles. They conclude that infiltration is indeed successful, can affect influence/popularity scores and possibly impact the social network as bots can manipulate trending topics during political and social campaigns.

Social impact of bots. In a recent work [9] authors devised a non-infiltrating honeypot experiment to study the impact of bots on content popularity. In [6], Edwards *et al.* highlight a positive view on the existence of bots on social media by studying the differences in perceptions of the quality of communication for a human agent and a bot agent on Twitter. They find that Twitter bots can be viewed as credible, attractive, competent in communication, and interactive.

Bot detection. In [21], Yan studied if an automated Turing test such as the CAPTCHA is sufficient to verify that an entity behind a computer is a human or an algorithm. The study concludes that CAPTCHA, apart from being inappropriate for some usability concerns, is insufficient to discern humans from bots. In a comprehensive work [3], Chu *et al.* distinguish and identify Twitter accounts operated by three entities: humans, cyborgs and bots. The authors make this classification by observing the differences among the three entities in terms of tweeting behaviour, tweet content and account properties. Using $1,000$ training samples the authors devised a system that classified their subset of the Twitter population into $5:4:1$ proportions for human:cyborg:bot, respectively. However, they neither provide an API for evaluation nor share datasets. The importance of bot detection on social media has recently gained momentum due to the rapid rise of bots. DARPA organised a Twitter bot challenge in 2016 [19] to detect influence bots – bots that illicitly shape topical discussions on Twitter to serve the purposes of their masters. DARPA provided $7,038$ accounts as ground truth labels that they knew about to the six teams who participated. The report concludes that detection of evolving influence bots requires carefully designed workflow.

However, as mentioned earlier most of the techniques neither expose their datasets nor their tools, which makes evaluation tough. To the best of our knowledge there is only one freely available and useable research tool, BOTORNOT [5], that detects bots on Twitter. The tool applies a Random Forests classifier and uses six groups of features to classify accounts as 'bots' or 'humans'. The model is trained using a list of social bots identified in [15] and a dataset from the Twitter Search API of 200 most recent tweets of these bots and 100 most recent tweets mentioning these bots. Apart from using a Random Forests classifier and a similar feature-set, we use raw historical data to cater for evolution of agents and

stealthier agents. We use a dataset partitioned into four popularity bands representing Twitter population at a more granular level, as agents differ according to the popularity and purpose of their creation and presence. We use 14 novel features from a set of total 21 attributes. Furthermore, we employ account categorisation in the preprocessed and partitioned datasets, and perform ablation tests to identify distinct group of features that are most effective for each popularity band (see §IV).

IV. Methodology

A tweet is formed of attributes written in JSON structure. Features we consider in this study are defined in Table I. While there are some features that have been used by previous studies [3], [5], a number of features that we consider in this work are used for the first time for this purpose to the best of our knowledge. These include: (*i*) *favourites-to-tweets ratio*, (*ii*) *lists per user*, (*iii*) *likes/favourites per tweet*, (*iv*) *retweets per tweet*, (*v*) *user replies*, (*vi*) 7 *activity source identity (or source type) categories*, (*vii*) *source count*, and (*viii*) *CDN content size*.

TABLE I
Features

Feature	Description and Hypotheses
Age of account	The age of the Twitter account in days. We assume humans have older accounts.
Favourites-to-tweets ratio	'Favourites' or 'likes' received for all user tweets. Humans tend to receive more 'likes' [10].
Lists per user	Lists subscribed to. Agents typically follow more lists to obtain lists of users to follow.
Followers-to-friends ratio	Previous research [3] shows that humans typically have this ratio close to 1.
User favourites	Tweets 'favourited' by a user. 'Liking' a post suggests an agreement, thus pointing to human-like behaviour.
Likes/favourites per tweet	'Favourites' received by a user. Humans typically receive more 'likes' [10], owing to content originality and topic diversity.
Retweets per tweet	'Retweets' received by a user. Humans typically receive more 'retweets' [10], owing to content originality and topic diversity.
User replies	Tweets replied to by a user. We assume humans engage in conversations with other users.
User tweets	User-generated tweets. Agents tweet more aggressively [10].
User retweets	User-generated retweets. Aggressive retweeting is a sign of automation [3].
Tweet frequency	Daily tweet frequency of a user. Agents often tweet much more often than humans per day [10].
URLs count	URLs are used to redirect traffic to elsewhere from Twitter platform. Presence of URLs within tweets suggests automation [3].
Activity source type	A 'source' is the endpoint from where a user posts tweets, denoted as S_n. We categorise this as: browser or web client (S_1), mobile device apps (S_2), social media management apps (S_3), social media scheduling and automation (S_4), marketing and brand promotion (S_5), news content web services (S_6), any other not part of the defined list (S_0). Humans mostly use S_1, S_2, and S_3; whereas agents mostly use S_4, S_5, and S_6.
Source count	The number of the endpoints used. We assume humans use more sources.
CDN content size	Content uploaded on Twitter. Agents on average tend to upload more content on Twitter [10].

We use the ***Stweeler***[11] platform [11] for collecting data, defining subsets, filtering data, calculating feature values and various other preprocessing, pre-analysis and classification tasks. For the purposes of this study we used a dataset collected in the month of April in 2016. We do not mention any keywords and collect everything offered by the Streaming API. On average the Streaming API provides between 2.5 and 3 million tweets a day.

Next, we partition the dataset.[12] into four user popularity bands to compensate for the variety of the purposes and activities associated with the popularity of the accounts. Note that we select accounts randomly for each band, and do not filter accounts based on any criteria (language or location). For example, an account with global recognition would not post malicious content. Most popular accounts are mostly legitimate, irrespective of being automated or human operated. Similarly, it is much more likely to find automated agents surreptitiously operating a less popular account. Moreover, we partition the dataset in *only* four popularity bands because these four groups are a sample that largely represents Twitter population where each metric follows Gaussian distribution, as detailed in a recent study [10]. The top two bands (Band_{10M} and Band_{1M}) are similar in their characteristics. These represent 0.04% and 0.61%, respectively, of the total partitioned accounts. Band_{1K} represents the bulk of Twitter – the most commonly found accounts representing users with ordinary popularity (approximately 94.40% of the total partitioned accounts in our dataset). Band_{100K} bridges the gap between the most popular and least popular ones, representing approximately 4.93% of the total partitioned accounts.

1) **Band**$_{10M}$: a subset of Twitter users with the highest number of followers, *i.e.,* $\geq$ $9M$ followers. These include 50 (24 agent and 26 human accounts)[13] most popular users that hold celebrity status (humans) and are globally renowned (agents). Such accounts (*e.g.,* CNN, BBC-World, NatGeo) post high quality content.
2) **Band**$_{1M}$: a subset of Twitter users that have between $0.9M$ and $1.1M$ followers. These include 746 (295 agent and 450 human accounts, and 1 tie) very popular users (*e.g.,* AlArabiya, dominos, pcgamer) that are close to attaining the celebrity status and global recognition.
3) **Band**$_{100K}$: a subset of $1,447$ user accounts (707 agent and 740 human accounts, and 1 tie) that include users having between $90K$ and $110K$ followers. These accounts (*e.g.,* Amtrak, BoobsVIP, CBSNewYork) that show high activity and have a bigger accumulated impact on Twitter.
4) **Band**$_{1K}$: a subset of $1,293$ Twitter users (499 agent and 794 human accounts) having between $0.9K$ and $1.1K$ followers which represent most commonly found (ordinary) users that form the bulk of the social graph (*e.g.,* ALTLENE_bot, hope_bot, Taiwan_Agent, TickerReport).

As examples depict, not all agents have malicious intent, indulge in infiltration or are purposed to be spammers.

[11]***Stweeler***– https://github.com/zafargilani/stcs

[12]Datasets can be found here – https://goo.gl/SigsQB. Classifier is available as a part of ***Stweeler***.

[13]This division is based on the majority vote of 4 annotators – see §V for more detail.

TABLE II
SUMMARY OF THE TWITTER ACCOUNTS DATASET.

Band	Followers	Accounts	# Tweets
Band_{10M}	$\geq 9M$	50	150,336
Band_{1M}	$0.9M - 1.1M$	746	303,517
Band_{100K}	$90K - 110K$	1,447	230,577
Band_{1K}	$0.9K - 1.1K$	1,293	37,679
Total		3,536	722,109

We then ask human participants to perform a cognitive or *human annotation task* to identify agents and humans. Chosen annotators are trained computer scientists and active Twitter users. Each annotator is given the same lists divided into the aforementioned four popularity bands and human annotation task guidelines[14] that outline a set of account properties and rules. Note that from the accounts provided, human annotators discard all the accounts that are: (*i*) no longer available, (*ii*) suspended or deleted, or (*iii*) have annotation score (bot:human) tied at 2:2. The task was time-bounded which meant that only a subset of the total accounts were annotated. Table II summarises the data in each of the popularity bands that were successfully annotated.

Further to the human annotation task, we collect annotations for the four popularity bands from the latest implementation of BOTORNOT for evaluation and comparison purposes. We use BOTORNOT for comparison purposes because hardly any past research makes the API to their tool or their datasets available online.[15] Also, hardly any past work compares other detection or classification tools to their experiments. We use BOTORNOT's HTTP REST API, which returns a bot-likelihood score for each Twitter account. BOTORNOT does not assign labels as 'bot' or 'human', but a 50% threshold (inferred from BOTORNOT website) is set as the boundary between an account being a human account (i.e. $< 50\%$ likelihood) and an account being a bot account ($>= 50\%$ likelihood). Whenever BOTORNOT returns a bot-likelihood score of less than 50% we assign 'human' label to that account, otherwise we assign 'agent' label.

We assume that the human annotation task produces a dataset annotated with the labels that are the closest approximations of the "ground truth" labels, since the latter are, in general, unavailable (see the discussion in §V). Furthermore, we use the agreement between the human annotators to benchmark the performance of the automated agent classification system.

We then calculate statistics for various features listed in Table I, and use a Random Forests classifier to perform three sets of experiments. First, we run a 5-fold cross-validation experiment in which we use 4 folds to train and 1 fold to test the classifier in each of the runs, with each fold containing subsets of all popularity bands, and report the results averaged across all 5 runs. Second, we report the results on the data originating with each of the popularity bands in particular. Third, we test how generalisable the features are, and for that we train the classifier using sets of 3 popularity bands and test it on 1 remaining popularity band in each of the runs.

We perform ablation tests: we start with the full feature-set and then remove features one by one in order to detect the minimal optimal feature combination that yields the best results on the task. Features that show up most often in the best performing feature splits in our experiments include *followers-to-friends ratio*, *user retweets*, *tweet frequency* and *URLs count*.

We finally obtain the classified datasets as well as the best features and their respective feature splits. Results of the annotation task and agent classification are presented in §V and §VI, respectively.

V. HUMAN ANNOTATION TASK

The annotation task fulfils two goals: first, it is used to derive the ground truth labels for the machine learning experiments presented in §VI. We cannot reliably use the information provided by the Twitter users on their accounts. Depending on the goals of a Twitter account operated by an agent, it may or may not self-identify as such: *e.g.,* if the goal is to spread false information and malicious content, the agent may pretend to be a human.

Second, human annotation task helps us estimate how accurately humans can identify agents on Twitter. This provides a very useful point of comparison for the machine learning experiments presented in §VI. The ultimate goal is to implement an automated tool for agent classification on Twitter that would perform comparably to humans, but it might be unrealistic to expect it to outperform humans. We will therefore compare the performance of the classifier presented in §VI to the inter-annotator agreement.

Human annotators have been provided with specific instructions (see §IV for more details). Twitter data within each popularity band has been independently annotated by 4 annotators. Each account is marked as either human or agent, and final ground truth labels are used (in the following machine learning experiments) *iff* majority vote holds between all annotators. Table III reports the average pairwise inter-annotator agreement across all popularity bands. In addition, we report average annotators' agreement with the final annotation, and average agreement of the annotators with the labels assigned by BOTORNOT (BON) [5]. The inter-annotator agreement in Table III is reported on the scale from 0% to 100%, with 0% showing lack of agreement and 100% being perfect agreement.

Table IV reports Cohen's *kappa* (κ) coefficient widely used in annotation experiments for assessing how reliable the annotators' judgements are, or determining "the degree, significance, and sampling stability of their agreement" [4]. This coefficient takes into account the observed agreement between the annotators p_o as well as the agreement that is expected by chance p_c, that is estimated by finding the joint probabilities of the marginals. The κ coefficient is calculated as follows:

[14]Human annotation task – `http://bit.ly/2cH0YvA`

[15]At the time of writing BOTORNOT does not open access to the details of their feature extraction and classifier, but provides an API – `https://botometer.iuni.iu.edu/`

TABLE III
AVERAGE INTER-ANNOTATOR AGREEMENT (%-AGE).

Ann	**Band**$_{10M}$	**Band**$_{1M}$	**Band**$_{100K}$	**Band**$_{1K}$
An$_1$	94.50	82.14	73.15	91.32
An$_2$	95.50	79.46	72.02	89.75
An$_3$	95.50	75.63	68.32	86.87
An$_4$	90.50	79.69	70.88	90.72
Avg	95.58	80.65	73.00	90.40
Final	**96.00**	**86.32**	**80.66**	**93.35**
BON	46.00	58.58	42.98	44.00

TABLE IV
AVERAGE COHEN'S κ.

Ann	**Band**$_{10M}$	**Band**$_{1M}$	**Band**$_{100K}$	**Band**$_{1K}$
An$_1$	89.00	63.26	46.37	81.68
An$_2$	90.93	57.90	44.21	77.99
An$_3$	90.93	50.41	36.69	72.17
An$_4$	80.86	58.03	41.71	80.14
Avg	85.15	60.27	46.05	79.58
Final	**91.96**	**71.76**	**61.28**	**85.91**
BON	-8.69	01.90	-14.46	-14.70

$$\kappa = \frac{p_o - p_c}{1 - p_c} \tag{1}$$

Following interpretation of κ values provided by [13], we conclude that annotators in our experiment achieve moderate ($\kappa \in [0.41 - 0.60]$ for band$_{100K}$) to substantial ($\kappa \in [0.61 - 0.80]$ for band$_{1K}$ and band$_{1M}$) to almost perfect ($\kappa \in [0.81 - 0.99]$ for band$_{10M}$) agreement which can be considered reliable in all cases. It is also worth noting that agreement of BOTORNOT with human annotators ranges from less than chance ($\kappa < 0.00$ for band$_{1K}$, band$_{100K}$ and band$_{10M}$) to slight ($\kappa \in [0.01 - 0.20]$ for band$_{1M}$) agreement only, which shows that human annotators almost always disagree with the labels assigned by BOTORNOT.

Interestingly, we observe the highest disagreement for the band$_{100K}$. Less particular properties within this band make these accounts similar to each other: *e.g.,* the annotators reported that a number of accounts within this band seemed to be initially agent-operated but were personalised later as human users started actively using them, and vice versa. This interesting phenomenon is worth exploring in the future.

Based on the results of the annotation task we conclude that: (*i*) The annotators mostly agree when they assign labels to the Twitter accounts, and the annotation can be considered reliable for all bands. (*ii*) BOTORNOT does not perform well on the given data and shows considerably large disagreement with human annotators' votes. (*iii*) We set the human annotation-based benchmark for the machine learning experiments reported in §VI at 87.42, or at the average observed agreement of the annotators with the final labels on the whole dataset spanning all four popularity bands.

VI. CLASSIFYING AGENTS AND HUMANS

We approach agent classification on Twitter as a binary classification task. Some previous research [3] distinguished between bots, humans and *cyborgs* – accounts that are partly operated by humans and also include automation, thus having properties of both bots and humans. In this work, we choose to perform binary classification distinguishing between agents and humans only, because accounts that consistently involve automation (*e.g.,* automated tweeting) should be characterised as automated accounts. As we noted in §I, our primary goal is to present a thorough methodological mechanism that allows identification of Twitter accounts as agents and humans using supervised classification.

TABLE V
DATASET BENCHMARKS.

Band	**Majority baseline**	**Human agreement**	**BON**
Band$_{10M}$	52.00	96.00	46.00
Band$_{1M}$	60.50	86.32	58.58
Band$_{100K}$	51.24	80.66	42.98
Band$_{1K}$	61.41	93.35	44.00
Total	**56.28**	**89.08**	**47.89**

We apply Random Forests classifier implemented using `scikit- learn` toolkit[16] and 100 decision tree estimators. We first define the benchmarks against which the automated account classification system is evaluated. The lower bound is set as the majority class distribution in the data, which for all popularity bands is equal to the proportion of accounts that belong to humans. In other words, if the automated account classification system always "guesses" that an account belongs to a human, then it will perform at the majority class baseline level. Next, we use the average observed inter-annotator agreement between each of the annotators and the final annotation, which tells us how well humans perform on this task as it may be unrealistic to expect an automated system to outperform humans (see §V). Finally, we also include the average agreement between the annotators and labels assigned by BOTORNOT. Table V reports these estimates for each of the popularity bands as well as the average across all data points in the whole dataset.

We perform three types of machine learning experiments (see §VI-A, VI-B, and VI-C) aimed at detecting how informative and generalisable features, that we overview in §IV, are for this task. For each of the experiments, we report accuracy of classification (`Acc`) which shows the proportion of agent and human accounts that the classifier identifies correctly, and precision (`P`), recall (`R`) and $\texttt{F}_1$ measures on the class of agents which show classifier's performance in identifying agents specifically.

A. Classifying agents by training and testing on all bands with 5-fold cross-validation

In the first experiment, we apply 5-fold cross-validation: we split the data into 5 non-overlapping folds, each containing approximately equal proportion of data points from each of the popularity bands, as well as having similar distribution of human and agent accounts. We then run the classifier over the folds, using each of the 5 folds as a test set once and training

[16]scikit-learn toolkit – `http://scikit-learn.org/`

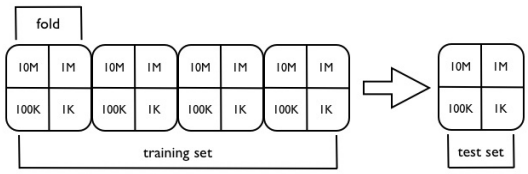

Fig. 1. Classifying agents by training and testing on all bands with 5-fold cross-validation.

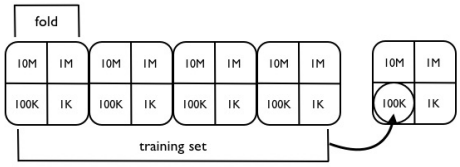

Fig. 2. Classifying agents by training on all and testing on specific bands with 5-fold cross-validation.

the classifier on the other 4 folds for each of the runs. Figure 1 illustrates this experiment. The first row (***Total***) of Table VI reports the results obtained with the best-performing feature-sets. This type of test enables us to determine the general accuracy of the classifier.

We run ablation tests to detect the most optimal feature-set – the minimal feature-set that yields the best accuracy. Ablation tests show that among the total of 21 features that we use in this work 12 features score among the most informative features across all 5 folds in the cross-validation experiment. These include *user replies*, *retweets per tweet*, *tweet frequency*, *age of account*, *followers-to-friends ratio*, *favourites-to-tweet ratio*, *URLs count*, and S_1, S_2, S_3, S_5, S_0. Note that human annotators also mentioned similar characteristics as strong indicators. A group of 6 other features score well for 4 out of 5 folds. These include *user tweets*, *user retweets*, *user favourites*, *likes/favourites per tweet*, *lists per user*, and S_4. Based on these results, we conclude that features that represent content dissemination (frequently tweeting, retweeting, posting URLs with tweets) and user engagement (following, receiving likes, receiving retweets, subscribing to lists) are overall the strongest predictors of automation.

Interestingly, *activity source* count and *CDN content size* considered in this experiment do not score as frequently among the most discriminative features on the data that combines all popularity bands. The annotators noted that the use of the Twitter API or automated activity source was a strong indicator of an automated behaviour on Twitter. This is confirmed by the nature or type of the activity sources (S_1 = browser, S_2 = mobile apps, S_3 = management, S_5 = marketing, and S_0 = all other services), all of which are strong indicators of automation.

B. Classifying agents by training on all and testing on specific bands with 5-fold cross-validation

In the second experiment, we train our classifier using the same 5 training folds containing data from all popularity bands, but report the results and run the ablation tests on the subsets of the test data that belong to each of the 4 popularity bands separately. Figure 2 describes the design of this experiment. In essence, the classifier is trained on the features that describe accounts from all 4 bands, but is then applied to the test data from one particular popularity band.[17] This experiment helps us discriminate between the results obtained on the data points originating within different popularity bands. Table VI reports the results.

[17]Note that the data in the training and test sets is non-overlapping as before: *i.e.,* each of the 5 test folds contains a different 20% of the data, with the rest being used for training.

TABLE VI
MACHINE LEARNING EXPERIMENTS RESULTS.

Band	**Acc**	$\mathbf{P}_{agents}$	$\mathbf{R}_{agents}$	$\mathbf{F1}_{agents}$
Total	86.44	85.40	82.20	83.60
Band_{10M}	100.00	100.00	100.00	100.00
Band_{1M}	91.76	90.60	88.00	89.40
Band_{100K}	85.70	85.60	85.40	85.60
Band_{1K}	88.25	87.80	80.80	84.00

We note that the performance follows similar trends as we report for the human annotation experiments (see Table III and Table IV): the classifier performs the best on band_{10M} and the worst on band_{100K}, whereas we also noted that human annotators reach highest agreement on band_{10M} and lowest on band_{100K}. Interestingly, when we train the classifier on the data from all popularity bands and measure its performance on specific bands, the classifier's accuracy on band_{10M}, band_{1M} and band_{100K} is above human agreement, and closely approaches human agreement on band_{1K} (*cf.* Table VI and Table V). The most informative features include *retweets per tweet*, *lists per user*, *tweet frequency*, *CDN content size*, and S_2, S_4. We note that features such as *age of account*, *follower-to-friend ratio*, *favourites-to-tweet ratio*, and *URLs count* that were informative when we combined data from all popularity bands, are not discriminative when we look at the popularity bands separately. On the contrary, features such as *lists per user*, *CDN content size* and S_4 = automation services, were not informative for combined data but are discriminative upon observing popularity bands separately.

C. Cross-band experiments

We then test how well the system generalises across the popularity bands with respect to the features used. For that, for each popularity band we train a classifier on the data from other 3 popularity bands and apply it to the particular band (see Figure 3). The experimental design is described in Figure 3, and the results are reported in Table VII.

We note that the classifier performance is consistently high for all bands, reaching the highest for band_{10M}. This effect might also be due to the size of the training and test sets: the ratio is the highest for band_{10M} with $3,486$ training and 50 test cases, and the lowest for band_{100K} with $2,089$ training and $1,447$ test cases. Nevertheless, we note that the performance on all bands is stable, with the accuracy being significantly above the majority class baseline as well as BOTORNOT

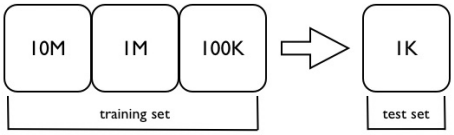

Fig. 3. Cross-band experiments.

TABLE VII
CROSS-BAND EXPERIMENTS RESULTS.

Band	**Acc**	$\mathbf{P}_{agents}$	$\mathbf{R}_{agents}$	$\mathbf{F1}_{agents}$
Band_{10M}	90.00	83.00	100.00	91.00
Band_{1M}	86.73	83.00	82.00	83.00
Band_{100K}	81.65	82.00	80.00	81.00
Band_{1K}	84.17	87.00	70.00	77.00

performance (see Table V).

We also note the effect of the training data size on generalisability of the feature-set itself: the largest training set for band_{10M} allows the classifier to achieve an accuracy of 90.00 using only 7 features (*user replies*, *follower-to-friend ratio*, *tweet frequency*, *favourites-to-tweet ratio*, and S_4 = automation services, S_5 = marketing, S_6 = news content web services), while the smallest training set for band_{100K} allows the classifier to achieve an accuracy of 81.65 relying on 16 out of the total of 21 features. The features that are most informative across all the bands include *age of account*, *user replies*, *retweets per tweet*, *tweet frequency*, *favourites-to-tweets ratio*, and S_4 = automation services, S_5 = marketing, S_6 = news content web services. We conclude that this set represents the most generalisable features that are quite independent of the type of account (*i.e.,* popularity level). We also note that they are in general consistent with the features that score well in other experiments, as well as the account properties that human annotators considered important when making their decisions (see §V).

D. Hypotheses testing

Finally, we check and report whether the features that we use in this work comply with our original hypotheses. For instance, we have assumed that agents tweet more aggressively than humans do and, thus, an average tweet frequency should be significantly higher for agent accounts than for human ones. In the last set of experiments, we apply t-test to the features for the humans and agents within each band and report: (1) whether the difference is statistically significant, and (2) whether it supports our original hypotheses in terms of the sign of the difference between the means.

Table VIII reports the results: we use + where the values for agent accounts are higher than those for human accounts, and − when human accounts have higher values; ** denotes statistical significance at 99% confidence level and * at 95% confidence level.

We note that these results are generally in accordance with our assumptions and also corroborate annotators' feedback as well as classification results: *e.g., tweet frequency*, S_2 = mobile apps, S_4 = automation services, S_5 = marketing, S_0 = all other services, and *source count* show the highest statistical significance overall. To summarise, there are several trends worth noting:

TABLE VIII
FEATURE SIGNIFICANCE.

Feature	$10M$	$1M$	$100K$	$1K$	All
Age of account	$+^{**}$	+	-	$-^{**}$	-
Favourites-to-tweets ratio	$-^{*}$	+	-	$-^{**}$	-
Lists per user	$-^{*}$	$+^{**}$	$+^{*}$	$+^{**}$	-
Followers-to-friends ratio	+	+	-	$+^{**}$	+
User favourites	+	-	$-^{**}$	-	$-^{**}$
Likes/favourites per tweet	$-^{**}$	N/A	N/A	N/A	$-^{**}$
Retweets per tweet	$-^{**}$	N/A	N/A	N/A	$-^{**}$
User replies	-	+	+	+	$+^{**}$
User tweets	-	$+^{*}$	$+^{**}$	$+^{**}$	+
User retweets	-	$+^{**}$	$+^{**}$	$+^{**}$	$+^{**}$
Tweet frequency	+	$+^{**}$	$+^{**}$	$+^{**}$	$+^{**}$
URLs count	+	+	$+^{**}$	$+^{**}$	$+^{**}$
S_1 = browser	+	+	-	-	-
S_2 = mobile apps	$-^{**}$	$-^{**}$	$-^{**}$	$-^{**}$	$-^{**}$
S_3 = OSN management	$+^{*}$	$+^{**}$	-	-	$+^{**}$
S_4 = automation	$+^{**}$	$+^{**}$	$+^{**}$	$+^{**}$	$+^{**}$
S_5 = marketing	$+^{*}$	$+^{*}$	$+^{**}$	$+^{**}$	$+^{**}$
S_6 = news content	$+^{*}$	+	+	N/A	$+^{*}$
S_0 = all other	$+^{*}$	$+^{**}$	$+^{**}$	$+^{**}$	$+^{**}$
Source count	$+^{**}$	$+^{**}$	$+^{**}$	$+^{**}$	$+^{**}$
CDN content size	+	+	$+^{**}$	$+^{**}$	$+^{*}$

- *Age of account* is a good predictor at the extreme ends of the popularity bands. At the same time, within the high popularity bands the agent accounts (*e.g.,* those of news agencies) are significantly older than human accounts (*e.g.,* those of celebrities). At the lower popularity levels, the difference is exactly the opposite, with the human accounts being significantly older than agent accounts.
- Humans in the high popularity band_{10M} follow significantly more lists than agents, while within the other bands agents join significantly more lists.
- Humans in the high popularity band_{10M} post more replies, and also tweet and retweet more than agents. Within the other popularity bands the trends change to exactly the opposite.
- The *number of URLs* posted, as well as the *CDN content size*, are higher for agents across all popularity bands, but the difference becomes statistically significant for band_{100K} and band_{1K}.
- S_2 = mobile app usage is significantly higher for humans than agents in all popularity bands.
- Usage of S_4 = automation services, S_5 = marketing and S_0 = all other services is significantly higher for agents than humans in all popularity bands.
- S_3 = OSN management seems to be employed by agents in band_{10M} and band_{1M}, while the opposite is true for band_{100K} and band_{1K}.
- The number of *source count* is significantly higher for agents in all popularity bands. This shows that within band_{10M} and band_{1M} humans post many URLs as well.

VII. Conclusion and Future Work

In this paper we developed and evaluated a thorough mechanism to reliably classify automated agents and human users on Twitter using a dataset divided into four popularity bands. We use a human annotation task to create ground truth labels, and verify the annotations using inter-annotator agreement among human annotators and BotOrNot (a bot detection research tool). We use a Random Forests classifier and perform three different machine learning experiments. Our classifier yields an accuracy that is on a par with human agreement for all four popularity bands. We also report on how different feature splits perform for different experiments. We note that 6 features show the highest statistical significance overall.

This work opens possibilities for related research in the future. A lot can be learned from topic analysis of the type of lists an account is following: *e.g.,* if the main goal of an agent is to expand its reach it can be assumed that the agent account would try to follow many different lists without particular topic coherence. This may also be useful as a feature.

Moreover, human annotation experiment (§V) shows that people pay attention to the content of the tweets: *e.g.,* human annotators cited the style and pattern of the tweets as strong indicators of agent-operated accounts, and also noted that abundance of promotional and depersonalised content strongly suggested that the account was operated by an automated agent. In this work, *URLs count* was used as one of the features to analyse the tweet content, with the higher number of URLs suggesting promotional and depersonalised content. Future research can focus on content analysis using NLP techniques to distinguish between the two entities. Another line of work can explore the provenance of social *botnets*, and ask if least popular Twitter accounts (having minimum activity) are being used to artificially inflate another account's popularity.

Acknowledgements: This work is funded by EU Metrics project (Grant EC607728).

References

[1] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, volume 6, page 12, 2010.

[2] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu. The socialbot network: When bots socialize for fame and money. In *Proceedings of the 27th Annual Computer Security Applications Conference*, ACSAC '11, pages 93–102, New York, NY, USA, 2011. ACM.

[3] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Who is tweeting on twitter: Human, bot, or cyborg? In *Proceedings of the 26th Annual Computer Security Applications Conference*, ACSAC '10, pages 21–30, New York, NY, USA, 2010. ACM.

[4] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[5] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, pages 273–274, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

[6] C. Edwards, A. Edwards, P. R. Spence, and A. K. Shelton. Is that a bot running the social media feed? testing the differences in perceptions of communication quality for a human agent and a bot agent on twitter. *Computers in Human Behavior*, 33:372–376, 2014.

[7] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. The rise of social bots. *Commun. ACM*, 59(7):96–104, June 2016.

[8] C. Freitas, F. Benevenuto, S. Ghosh, and A. Veloso. Reverse engineering socialbot infiltration strategies in twitter. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ASONAM '15, pages 25–32, New York, NY, USA, 2015. ACM.

[9] Z. Gilani, R. Farahbakhsh, and J. Crowcroft. Do bots impact twitter activity? In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 781–782, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.

[10] Z. Gilani, R. Farahbakhsh, G. Tyson, L. Wang, and J. Crowcroft. An in-depth characterisation of bots and humans on twitter. *arXiv preprint arXiv:1704.01508*, 2017.

[11] Z. Gilani, L. Wang, J. Crowcroft, M. Almeida, and R. Farahbakhsh. Stweeler: A framework for twitter bot analysis. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, pages 37–38, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

[12] B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about twitter. In *Proceedings of the First Workshop on Online Social Networks*, WOSN '08, pages 19–24, New York, NY, USA, 2008. ACM.

[13] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[14] K. Lee, J. Caverlee, and S. Webb. Uncovering social spammers: Social honeypots + machine learning. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 435–442, New York, NY, USA, 2010. ACM.

[15] K. Lee, B. D. Eoff, and J. Caverlee. Seven months with the devils: A long-term study of content polluters on twitter. In *ICWSM*, 2011.

[16] S. Savage, A. Monroy-Hernandez, and T. Höllerer. Botivist: Calling volunteers to action using online bots. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, pages 813–822, New York, NY, USA, 2016. ACM.

[17] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and W. B. Dolan. A neural network approach to context-sensitive generation of conversational responses. In *HLT-NAACL*, pages 196–205. Association for Computational Linguistics, May–June 2015.

[18] G. Stringhini, C. Kruegel, and G. Vigna. Detecting spammers on social networks. In *Proceedings of the 26th Annual Computer Security Applications Conference*, ACSAC '10, pages 1–9, New York, NY, USA, 2010. ACM.

[19] V. S. Subrahmanian, A. Azaria, S. Durst, V. Kagan, A. Galstyan, K. Lerman, L. Zhu, E. Ferrara, A. Flammini, and F. Menczer. The darpa twitter bot challenge. *Computer*, 49(6):38–46, June 2016.

[20] B. Viswanath, M. A. Bashir, M. Crovella, S. Guha, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Towards detecting anomalous user behavior in online social networks. In *Usenix Security*, volume 14, 2014.

[21] J. Yan. Bot, cyborg and automated turing test. In *International Workshop on Security Protocols*, pages 190–197. Springer, 2006.